NCBI Taxonomy Database for Prokaryotic Curation

Shobha Sharma and Conrad Schoch
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

Introduction

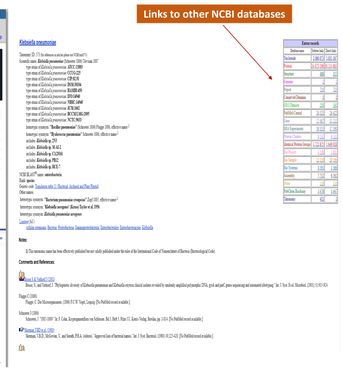
National Center for Biotechnology Information (NCBI) taxonomy database contains the names and phylogenetic lineages of more than **470,000** species with formal names that have molecular data in INSDC databases, about **19,000** of these are species level prokaryote names. Taxonomy database can be regarded as the central organizing hub for many of the resources at NCBI. Here are some of features of NCBI taxonomy database which may not be quite obvious to users.

NCBI Taxonomy Database

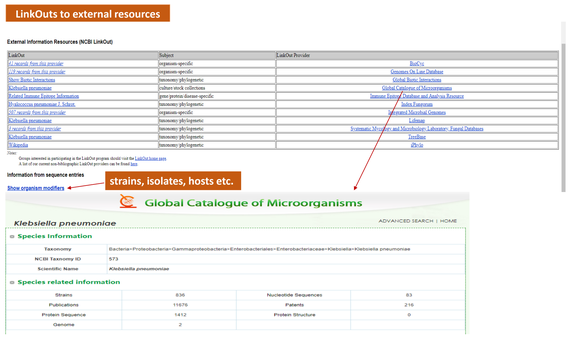
NCBI Taxonomy database is the standard nomenclature and classification repository for the International Nucleotide Sequence Database Collaboration (INSDC) which comprises GenBank, ENA and DDBJ databases. It is an entry point into the NCBI system for users who want to find all available information about a particular taxon.

https://www.ncbi.nlm.nih.gov/taxonomy

Links to other NCBI databases



LinkOuts to external resources



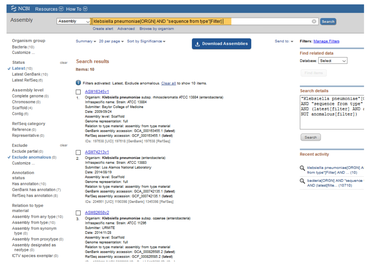
strains, isolates, hosts etc.

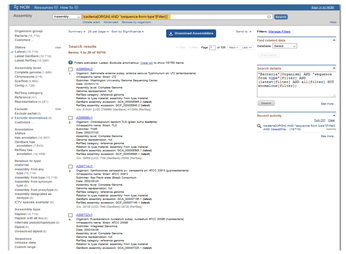
Type Material in the NCBI Taxonomy Database

Since 2013, GenBank curates type material (including synonym types) in the Taxonomy Database and uses it to flag sequences from type or synonym type in sequence records. Sequence from type is an important subset of GenBank.

Currently GenBank has over **10,700** prokaryotic genome assemblies from type strains.

You can search assemblies from type by using
"sequence from type"[Filter]





Correcting Prokaryotic Genomes Based on Average Nucleotide Identity (ANI)

All new prokaryote genomes submissions to GenBank go through average nucleotide identity (ANI) check to ensure asserted organism name is correct. ANI is also used to correct misidentified genomes that are already public in GenBank

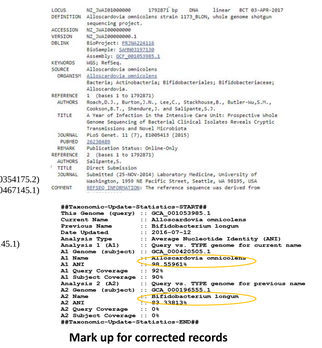
There are **~280,000** live bacterial assemblies in GenBank and **~10,000** assemblies from type stains

Submitted organism: Vibrio rotiferianus
Predicted organism: Vibrio parahaemolyticus
Submitted organism has type: Yes
Status: MISASSIGNED
Confidence: HIGH

ANI (QCov SCov) — **Organism (Accession)**

98.439 (82.5 90.0) Vibrio parahaemolyticus (GCA_001558495.2)
98.438 (82.4 91.3) Vibrio parahaemolyticus (GCA_010111015.1)
98.436 (82.3 91.2) Vibrio parahaemolyticus (GCA_001998785.1)
98.448 (81.7 91.9) Vibrio parahaemolyticus NBRC 12711 (GCA_000813305.1)
85.133 (68.9 76.0) Vibrio antiquarius (GCA_000024825.1)
85.105 (68.7 75.2) Vibrio diabolicus (GCA_001048675.1)
84.475 (65.0 70.5) Vibrio alginolyticus (GCA_001471275.2)
84.456 (65.0 71.0) Vibrio alginolyticus NBRC 15630 = ATCC 17749 (GCA_000354175.2)
84.452 (64.7 72.0) Vibrio alginolyticus NBRC 15630 = ATCC 17749 (GCA_000467145.1)
85.199 (63.7 73.9) Vibrio antiquarius (GCA_000152485.1)
82.803 (53.3 50.3) Vibrio harveyi (GCA_001525585.2)
82.542 (52.8 51.5) Vibrio harveyi (GCA_001471575.2)
82.603 (52.1 50.2) Vibrio owensii LMG 25430 (GCA_000400325.1,
82.495 (52.0 52.3) Vibrio harveyi NBRC 15634 = ATCC 14126 (GCA_001591145.1)
82.318 (51.8 49.0) Vibrio jasicida (GCA_001625175.1)

Correcting new submissions



Mark up for corrected records

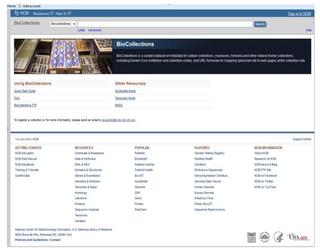
ANI Current Status

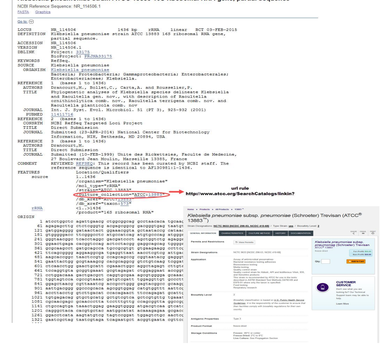
~10,000 type strains available for **~19,000** species
More types added as they are submitted
~800 species IDs have been corrected (public assemblies)
~2100 new submissions were corrected before they become public
~4,700 genome assemblies are currently misidentified
ANI process executed daily
Incorporated into PGAP for identification of problems before submission

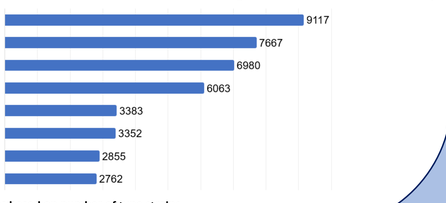
NCBI BioCollections Database

Curated dataset of metadata for culture collections, museums, herbaria and other natural history collections connected to sequence records in GenBank. Used to Support the "structured voucher" annotation in the sequence entries submitted to INSDC.

https://www.ncbi.nlm.nih.gov/biocollections







Top culture collections based on number of type strains

References

Federhen S. 2012. **The NCBI Taxonomy database**. Nucleic Acids Res. 40: D136-43.

Federhen S. 2015. **Type material in the NCBI Taxonomy Database**. Nucleic Acids Res. 43: D1086-98.

Sharma S., Ciufo S., Starchenko E., Darji D., Chumsky L., Karsch-Mizrachi I. and Schoch CL. 2018. **The NCBI Biocollections Database.** Database(Oxford) doi: 10.1093/database/baz057.

Ciufo S., Kannan S., Sharma S., Badretdin A., Clark K., Turner S. Brover S., Schoch CL., Kimchi A. and DiCuccio M. 2018. **Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI**. Int J Syst Evol Microbiol. 66:2386-2392. doi: 10.1099/ijsem.0.002809.